sas THE POWER TO KNOW

ETL

What it is and why it matters

ETL is a type of data integration that refers to the three steps (extract, transform, load) used to blend data from multiple sources. It's often used to build a data warehouse. During this process, data is taken (extracted) from a source system, converted (transformed) into a format that can be analyzed, and stored (loaded) into a data warehouse or other system. Extract, load, transform (ELT) is an alternate but related approach designed to push processing down to the database for improved performance.

Importance

Today's World

How It's Used

How It Works

ETL History

ETL gained popularity in the 1970s when organizations began using multiple data repositories, or databases, to store different types of business information. The need to integrate data that was spread across these databases grew quickly. ETL became the standard method for taking data from disparate sources and transforming it before loading it to a target source, or destination.

In the late 1980s and early 1990s, data warehouses came onto the scene. A distinct type of database, data warehouses provided integrated access to data from multiple systems – mainframe computers, minicomputers, personal computers and spreadsheets. But different departments often chose different ETL tools to use with different data warehouses. Coupled with mergers and acquisitions, many organizations wound up with several different ETL solutions that were not integrated.

Over time, the number of data formats, sources and systems has expanded tremendously. Extract, transform, load is now just one of several methods organizations use to collect, import and process data. ETL and ELT are both important parts of an organization’s broader data integration strategy.

Why ETL Is Important

Businesses have relied on the ETL process for many years to get a consolidated view of the data that drives better business decisions. Today, this method of integrating data from multiple systems and sources is still a core component of an organization's data integration toolbox.

ETL in Today's World

Today's fast-moving data (streaming data) can be captured and analyzed on the fly via streaming analytics. This approach presents the opportunity to act immediately, based on what's happening at a moment in time. But the historical view afforded by ETL puts data in context. In turn, organizations get a well-rounded understanding of the business over time. The two approaches need to work together.

7 Tips to Modernize Data Integration

Data integration has been

Benefits of a Single Customer View

This energy company

Data Integration Reimaginec

Instead of dying out, old technologies

Gartner

SAS: Leader in the 2017 Gartner Magic Quadrant

Gartner has

around for years, but it still plays a vital role in capturing, processing

stored customer data on different systems and in different formats

often end up coexisting with new ones. Today, data integration

positioned SAS as a Leader in the 2017 Gartner *Magic Quadrant for Data*

> The most successful organizations will have a clear and precise strategy in place that recognizes data integration as a fundamental cornerstone of their competitive differentiation.

–David Loshin, President of Knowledge Integrity Inc. *The New Data Integration Landscape: Moving Beyond Ad Hoc ETL to an Enterprise Data Integration Strategy*

Data Integration Software From SAS

Data integration software from SAS distributes integration tasks across any platform and virtually connects to any source or target data store.

Learn more about data integration software from SAS

How ETL Is Being Used

Core ETL and ELT tools work in tandem with other data integration tools, and with various other aspects of data management – such as data quality, data governance, virtualization and metadata. Popular uses today include:

ETL and Traditional Uses

ETL is a proven method that many organizations rely on every day – such as retailers who need to see sales data regularly, or health care providers looking for an accurate depiction of claims. ETL can combine and surface transaction data from a warehouse or other data store so that it's ready for business people to view

ETL With Big Data – Transformations and Adapters

Whoever gets the most data, wins. While that's not necessarily true, having easy access to a broad scope of data can give businesses a competitive edge. Today, businesses need access to all sorts of big data – from videos, social media, the in a format they can understand. ETL is also used to migrate data from legacy systems to modern systems with different data formats. It's often used to consolidate data from business mergers, and to collect and join data from external suppliers or partners.

ETL for Hadoop – and More

ETL has evolved to support integration across much more than traditional data warehouses. Advanced ETL tools can load and convert structured and unstructured data into Hadoop. These tools read and write multiple files in parallel from and to Hadoop, simplifying how data is merged into a common transformation process. Some solutions incorporate libraries of prebuilt ETL transformations for both the transaction and interaction data that run on Hadoop. ETL also supports integration across transactional systems, operational data stores, BI platforms, master data management (MDM) hubs and the cloud.

Internet of Things (IoT), server logs, spatial data, open or crowdsourced data, and more. ETL vendors frequently add new transformations to their tools to support these emerging requirements and new data sources. Adapters give access to a huge variety of data sources, and data integration tools interact with these adapters to extract and load data efficiently.

ETL and Self-Service Data Access

Self-service data preparation is a fast-growing trend that puts the power of accessing, blending and transforming data into the hands of business users and other nontechnical data professionals. Ad hoc in nature, this approach increases organizational agility and frees IT from the burden of provisioning data in different formats for business users. Less time is spent on data preparation and more time is spent on generating insights. Consequently, both business and IT data professionals can improve productivity, and organizations can scale up their use of data to make better decisions.

ETL and Data Quality

ETL and other data integration software tools – used for data cleansing, profiling and auditing – ensure that data is trustworthy. ETL tools integrate with data quality tools, and ETL vendors incorporate related tools within their solutions, such as those used for data mapping and data lineage.

ETL and Metadata

Metadata helps us understand the lineage of data (where it comes from) and its impact on other data assets in the organization. As data architectures become more complex, it's important to track how the different data elements in your organization are used and related. For example, if you add a Twitter account name to your customer database, you'll need to know what will be affected, such as ETL jobs, applications or reports.

SAS® Data Management in Action

With SAS Data Management, you can take advantage of huge volumes of data – for example, customer data from Twitter feeds – to get insights like never before. Matthew Magne explains how SAS can stream Twitter data into a data lake, cleanse and profile the data, then reveal which customers are most likely to leave. In turn, you can create a plan to retain them.

How It Works

ETL is closely related to a number of other data integration functions, processes and techniques. Understanding these provides a clearer view of how ETL works.

SQL	Structured query language is the most common method of accessing and transforming data within a database.
Transformations, business rules and adapters	After extracting data, ETL uses business rules to transform the data into new formats. The transformed data is then loaded into the target.
Data mapping	Data mapping is part of the transformation process. Mapping provides detailed instructions to an application about how to get the data it needs to process. It also describes which source field maps to which destination field. For example, the third attribute from a data feed of website activity might be the user name, the fourth might be the time stamp of when that activity happened, and the fifth might be the product that the user clicked on. An application or ETL process using that data would have to map these same fields or attributes from the source system (i.e., the website activity data feed) into the format required by the destination system. If the destination system was a customer relationship management system, it might store the user name first and the time stamp fifth; it might not store the selected product at all. In this case, a transformation to format the date in the expected format (and in the right order), might happen in between the time the data is read from the source and written to the target.
Scripts	ETL is a method of automating the scripts (set of instructions) that run behind the scenes to move and transform data. Before ETL, scripts were written individually in C or COBOL to transfer data between specific systems. This resulted in multiple databases running numerous scripts. Early ETL tools ran on mainframes as a batch process. ETL

	later migrated to UNIX and PC platforms. Organizations today still use both scripts and programmatic data movement methods. .
ETL versus ELT	In the beginning, there was ETL. Later, organizations added ELT, a complementary method. ELT extracts data from a source system, loads it into a destination system and then uses the processing power of the source system to conduct the transformations. This speeds data processing because it happens where the data lives.
Data quality	Before data is integrated, a staging area is often created where data can be cleansed, data values can be standardized (NC and North Carolina, Mister and Mr., or Matt and Matthew), addresses can be verified and duplicates can be removed. Many solutions are still standalone, but data quality procedures can now be run as one of the transformations in the data integration process.
Scheduling and processing	ETL tools and technologies can provide either batch scheduling or real-time capabilities. They can also process data at high volumes in the server, or they can push down processing to the database level. This approach of processing in a database as opposed to a specialized engine avoids data duplication and prevents the need to use extra capacity on the database platform.
Batch processing	ETL usually refers to a batch process of moving huge volumes of data between two systems during what’s called a “batch window.” During this set period of time – say between noon and 1 p.m. – no actions can happen to either the source or target system as data is synchronized. Most banks do a nightly batch process to resolve transactions that occur throughout the day.
Web services	Web services are an internet-based method of providing data or functionality to various applications in near-real time. This method simplifies data integration processes and can deliver more value from data, faster. For example, let’s say a customer contacts your call center. You could create a web service that returns the complete customer profile with a subsecond response time simply by passing a phone number

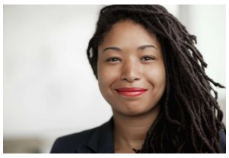
	to a web service that extracts the data from multiple sources or an MDM hub. With richer knowledge of the customer, the customer service rep can make better decisions about how to interact with the customer.
Master data management	MDM is the process of pulling data together to create a single view of the data across multiple sources. It includes both ETL and data integration capabilities to blend the data together and create a “golden record” or “best record.”
Data virtualization	Virtualization is an agile method of blending data together to create a virtual view of data without moving it. Data virtualization differs from ETL, because even though mapping and joining data still occurs, there is no need for a physical staging table to store the results. That’s because the view is often stored in memory and cached to improve performance. Some data virtualization solutions, like SAS Federation Server, provide dynamic data masking, randomization and hashing functions to protect sensitive data from specific roles or groups. SAS also provides on-demand data quality while the view is generated.
Event stream processing and ETL	When the speed of data increases to millions of events per second, event stream processing can be used to monitor streams of data, process the data streams and help make more timely decisions. An example in the energy space is using predictive analytics on streams of data to detect when a submersible pump is in need of repair to reduce both downtime and the scope and size of damage to the pump.

Read More About This Topic

Key questions to kick off your data analytics

5 ways to become data-driven

GDPR and AI: Friends, foes or something in between?



Personal data: Getting it right with GDPR

Leave a message